Bayesian Model Search and Multilevel Inference for SNP Association Studies

Melanie A. Wilson[1], Edwin S. Iversen[1], Merlise A. Clyde[1], Scott C. Schmidler[1]
and Joellen M. Schildkraut[2]
Departments of Statistical Science[1] and Community and Family Medicine[2],
Duke University, Durham, NC; 27713.
Correspondence: Melanie Wilson, E-mail: maw27@stat.duke.edu Phone: 919-684-4558.

October 13, 2008

Abstract

Technological advances in genotyping have given rise to hypothesis based association studies of increasing scope. In particular, candidate pathway studies of single nucleotide polymorphisms (SNPs) have all but replaced studies focusing on a small number of variants in a handful of genes. As a result, the scientific hypotheses addressed by these studies have become more complex and more difficult to address using existing analytic methodologies. Obstacles to analysis include inference in the face of multiple comparisons, complications arising from correlations among the SNPs and choice of their genetic parametrization. Here we describe a Bayesian model search strategy applied to penalized logistic regression that searches the space of genetic markers, and over the genetic parametrization of each, in a computationally efficient manner. This technique allows one to estimate multilevel posterior probabilities and Bayes Factors at the global, gene and SNP level. Using an ongoing study as an example, we describe how simulated data sets based on a study can be used both to identify optimal hyper–parameters for the prior distribution and to characterize the method's statistical power to identify associations given a choice of hyper–parameters. We conclude by describing an analysis of the data set upon which the simulations were based. The model search algorithm, functions to summarize its output and convergence diagnostics are implemented in the freely available `R` package, `MISA` (Multilevel Inference for SNP Associations).

1 Introduction

Recent advances in genotyping technology have resulted in a dramatic change in the way genetic association studies are conducted. Instead of investigating only a handful of variants within the most interesting genes, researchers are conducting candidate–gene or genome–wide studies that may encompass hundreds to a million or more genetic variants, often, single nucleotide polymorphisms (SNPs). As a result, there is a growing demand for statistical methods that are better suited to the scale and nature of contemporary association studies.

In this paper, we focus on intermediate throughput association studies, where the outcome of interest is a binary phenotype such as disease status and where the genetic markers have been chosen to capture variation in a set of related genes, such as those involved in a specific biochemical pathway. We define pathway as a set of genes thought to be active in certain circumstances. Statistical analysis of data from such studies will typically address two questions: 'To what extent do the data support an overall association between the pathway and outcome of interest?' and 'Which genes or variants are most likely to be driving this association?' We describe a modeling approach that addresses both of these questions.

Many analyzes of candidate pathway association data proceed by assuming conditional independence of the genetic markers given disease status. These methods have proved to be extremely useful given their simplicity, but have several significant drawbacks when used to quantify over–all or marker–by–marker association. These include issues with multiple comparisons, complications arising from correlations among the genetic markers, and choice of the parametrization of the genetic variables. Of these, the most attention has been aimed at the multiple testing issue. Numerous methods have been developed to determine if a SNP is 'noteworthy' given a corresponding p–value (Storey, 2002; Wacholder, 2004; Wakefield, 2007). Models for marker–specific test statistics have also been used to calculate measures of pathway–wide association (Tyrer et al., 2006). While these methods have been shown to be effective in controlling the number of false discoveries reported, they often make distributional assumptions on the p–values and/or on the markers themselves that are not always justifiable (Efron, 2007).

Prospective models for disease outcome given multivariate genetic marker data make weaker assumptions about the distribution of markers given disease and, importantly, provide measures of association that are adjusted for the remaining markers. Within this framework, Bayesian model

selection is an effective approach for identifying subsets of likely associated variables, for prioritizing them and for measuring overall association (Clyde and George, 2004). In this spirit, logic regression (Ruczinski et al., 2003) and its associated model search algorithms for penalized likelihoods have been applied successfully to association studies (Kooperberg and Ruczinski, 2004; Schwender and Ickstadt, 2007). This methodology provides estimates of the importance of individual SNPs and of their interactions with others in addition to providing a measure of global association. Our aims are similar, but we choose to concentrate on main effects under specified genetic models of association; hence, MISA's model space is a subset of logic regression's. In contrast, we focus on a fully Bayesian approach, giving considerable attention to the choice of prior and its influence on inference — especially as regards its effect on the properties of decision procedures based on its output — and on constructing formal measures of evidence for association at the global, gene and SNP level. Our desire is to provide a tool for summarizing the results of candidate gene/pathway studies and for determining which, if any, candidate SNPs, genes or pathways merit further investigation.

In this paper, we propose a Bayesian model search technique for logistic regression that searches over genetic markers, and over the genetic parametrization of each. Through posterior probabilities and Bayes factors, we are able to address hypotheses of association at multiple levels: pathway (global), genes and SNPs. In Section 2, we describe our approach for Multilevel Inference for SNP Associations (MISA); these methods are implemented in the freely available `R` package `MISA` . In Section 3, we apply our method to simulated data and demonstrate that `MISA` has desirable operating characteristics and demonstrate how the simulation study can be used to guide selection of the prior hyper-parameters given the study design. Using this prior choice, we analysis a candidate gene pathway study and highlight interpretation of the results. We conclude in Section 4 with a discussion.

2 Methods

2.1 Model Specification

We consider SNP association models with a binary phenotype, such as presence or absence of a disease. For $i = 1, \ldots, n$, we let D_i indicate the disease status of individual i, where $D_i = 1$ represents a disease case and $D_i = 0$ represents a control. For each individual, we measure S single

nucleotide polymorphisms (SNPs), where SNP s is either homozygous rare, $a_s a_s$, heterozygous, $a_s A_s$ or $A_s a_s$, or homozygous common, $A_s A_s$. In addition to the SNP data, for each individual we have a q dimensional vector $\mathbf{z}_i^T$ of design and potential confounding variables that will be included in all models; henceforth we refer to these as 'design' variables.

We use logistic regression models to relate disease status to the design variables and subsets of SNPs. We denote the collection of all possible models by $\mathcal{M}$. Individual models, denoted by $\mathcal{M}_{\boldsymbol{\gamma}}$, are specified by the S dimensional vector γ where γ_s indicates the inclusion of SNP s in model $\mathcal{M}_{\boldsymbol{\gamma}}$, and if included, specifies the SNP-specific genetic parametrization, where $\gamma_s = 0$ if $\text{SNP}_s \notin \mathcal{M}_{\boldsymbol{\gamma}}$, $\gamma_s = 1$ if $\text{SNP}_s \in \mathcal{M}_{\boldsymbol{\gamma}}$ with a log-additive parametrization, $\gamma_s = 2$ if $\text{SNP}_s \in \mathcal{M}_{\boldsymbol{\gamma}}$ with a dominant parametrization, and $\gamma_s = 3$ if $\text{SNP}_s \in \mathcal{M}_{\boldsymbol{\gamma}}$ with a recessive parametrization[1]. Under each of these genetic parametrization, SNP s may be coded using one degree of freedom. In particular, for the log-additive model the design variable representing SNP s is a numeric variable equal to the number of copies of the risk allele a_s. For the dominant model, we use an indicator variable of whether allele a_s is present (homozygous rare or heterozygous) and for the recessive model, an indicator variable of whether SNP s equals the homozygous rare genotype. For each individual, the logistic regression under model $\mathcal{M}_{\boldsymbol{\gamma}}$ is given by

$$\text{logit}(p(D_i = 1|\mathbf{z}_i, \mathbf{x}_{\boldsymbol{\gamma}_i}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}, \mathcal{M}_{\boldsymbol{\gamma}})) = \alpha_0 + \mathbf{z}_i^T \boldsymbol{\alpha} + \mathbf{x}_{\boldsymbol{\gamma}_i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}$$

where $\mathbf{x}_{\boldsymbol{\gamma}_i}$ represents the coding of SNPs included in model $\mathcal{M}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ is the vector of model specific parameters, $(\alpha_0, \boldsymbol{\alpha}^T, \boldsymbol{\beta}_{\boldsymbol{\gamma}}{}^T)$, with intercept α_0, vector of coefficients for the design variables $\boldsymbol{\alpha}$, and log odds ratios $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ for SNPs included in model $\mathcal{M}_{\boldsymbol{\gamma}}$.

2.2 Posterior Model Probabilities

Posterior model probabilities measure the degree to which the data support each in a set of competing models. The posterior model probability of any model $\mathcal{M}_{\boldsymbol{\gamma}}$ in the space of models $\mathcal{M}$ is expressed as

$$p(\mathcal{M}_{\boldsymbol{\gamma}} \mid D) = \frac{p(D \mid \mathcal{M}_{\boldsymbol{\gamma}}) p(\mathcal{M}_{\boldsymbol{\gamma}})}{\sum_{\mathcal{M}_{\boldsymbol{\gamma}} \in \mathcal{M}} p(D \mid \mathcal{M}_{\boldsymbol{\gamma}}) p(\mathcal{M}_{\boldsymbol{\gamma}})} \quad \text{for } \mathcal{M}_{\boldsymbol{\gamma}} \in \mathcal{M}$$

[1] In the case where no homozygous rare cases or controls are observed, we fixed the parametrization to be 'log–additive.'

where $p(D \mid \mathcal{M}_{\boldsymbol{\gamma}})$ is proportional to the (marginal) likelihood of model $\mathcal{M}_{\boldsymbol{\gamma}}$ obtained after integrating out model specific parameters $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$, and $p(\mathcal{M}_{\boldsymbol{\gamma}})$ is the prior probability of $\mathcal{M}_{\boldsymbol{\gamma}}$ which is discussed in more detail in Section 2.7.

We assume normal prior distributions for the coefficients $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ with a covariance matrix that is given by a constant $1/k$ times the inverse Fisher Information matrix. For logistic regression models, analytic expressions for $p(D \mid \mathcal{M}_{\boldsymbol{\gamma}})$ are not available and Laplace approximations or the Bayes Information Criterion are commonly used to approximate the marginal likelihood (Raftery, 1986; Clyde and George, 2004; Wakefield, 2007; Consortium, 2007). Using a Laplace approximation (see Appendix A), the posterior probability of model $\mathcal{M}_{\boldsymbol{\gamma}}$ takes the form of a penalized likelihood or deviance criterion

$$p(\mathcal{M}_{\boldsymbol{\gamma}} \mid D) \propto \exp\{-\frac{1}{2}[\mathrm{dev}(\mathcal{M}_{\boldsymbol{\gamma}}; D) + \mathrm{pen}(\mathcal{M}_{\boldsymbol{\gamma}})]\} \tag{1}$$

where $\mathrm{dev}(\mathcal{M}_{\boldsymbol{\gamma}}; D) = -2\log(p(D \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}, \mathcal{M}_{\boldsymbol{\gamma}}))$ is the model deviance and the penalty term, $\mathrm{pen}(\mathcal{M}_{\boldsymbol{\gamma}})$, encompasses a penalty on model size induced by the choice of k in the prior distribution on coefficients $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ and the prior distribution over models. Because we expect that there may be (many) small effect sizes, we calibrate the choice of k based on the Aikake information criterion, leading to

$$\mathrm{pen}(\mathcal{M}_{\boldsymbol{\gamma}}) = 2(1 + q + s_{\boldsymbol{\gamma}}) - 2\log(p(\mathcal{M}_{\boldsymbol{\gamma}})).$$

2.3 Missing Data

Missing SNP data are the norm rather than the exception in association studies. Removing subjects with missing SNP genotype data will typically eliminate too many observations resulting in an unnecessary loss of information. It is possible to take advantage of patterns in linkage disequilibrium to efficiently impute the missing genotypes given observed data. We use fastPHASE (Stephens et al., 2001) to sample haplotypes and missing genotypes (D^{miss}) given the observed unphased genotypes (D^{obs}). The posterior probabilities of models given the observed data are

$$\begin{aligned} p(\mathcal{M}_{\boldsymbol{\gamma}} \mid D^{\mathrm{obs}}) &\propto \int \exp\{-\frac{1}{2}[\mathrm{dev}(\mathcal{M}_{\boldsymbol{\gamma}}; D^{\mathrm{obs}}, D^{\mathrm{miss}}) + \mathrm{pen}(\mathcal{M}_{\boldsymbol{\gamma}})]\} p(D^{\mathrm{miss}} \mid D^{\mathrm{obs}}) dD^{\mathrm{miss}} \\ &\approx \frac{1}{M}\sum_{m=1}^{M} \exp\{-\frac{1}{2}[\mathrm{dev}(\mathcal{M}_{\boldsymbol{\gamma}}; D^{\mathrm{obs}}, D_m^{\mathrm{miss}}) + \mathrm{pen}(\mathcal{M}_{\boldsymbol{\gamma}})]\} \equiv \Psi(\mathcal{M}_{\boldsymbol{\gamma}}) \end{aligned} \tag{2}$$

where M is the number of imputed data sets and $\Psi(\mathcal{M}_{\boldsymbol{\gamma}})$ is an estimate of the un-normalized posterior model probability for model $\mathcal{M}_{\boldsymbol{\gamma}}$.

2.4 Model Search

Many current methods for analyzing data in SNP association studies assume that SNPs are conditionally independent given disease status which limits the space of models under consideration. In this framework, the model space $\mathfrak{M}$ would be comprised of S single SNP effect models where some *a priori* choice is made on each SNP-specific genetic parametrization. We aim to weaken these assumptions by expanding the model space to allow multiple SNPs and considering multiple parametrizations for each of the SNPs. Because the dimension of $\mathfrak{M}$ grows exponentially with S, we are unable to enumerate all models for S greater than 25–30. Stochastic variable selection algorithms (see Clyde and George, 2004, for a review) bypass the need to enumerate the space of models and permit calculation of posterior probabilities based on a sample of the most likely candidate models.

We develop a stochastic search algorithm based on the Evolutionary Monte Carlo (EMC) algorithm of Liang and Wong (2000). EMC is a combination of parallel tempering (Geyer, 1991) and a genetic algorithm (Holland, 1975), which samples models based on their "fitness". In our application, the fitness of the models is given by the fitness function $\psi(\mathfrak{M}_{\boldsymbol{\gamma}})$

$$\psi(\mathfrak{M}_{\boldsymbol{\gamma}}) = -\frac{1}{2}\left(\text{dev}(\mathfrak{M}_{\boldsymbol{\gamma}}) + \text{pen}(\mathfrak{M}_{\boldsymbol{\gamma}})\right)$$

which is equal to the log of the un-normalized posterior model probability from equation (1), or in the case of missing data

$$\psi(\mathfrak{M}_{\boldsymbol{\gamma}}) = \log(\Psi(\mathfrak{M}_{\boldsymbol{\gamma}}))$$

where $\Psi(\mathfrak{M}_{\boldsymbol{\gamma}}))$ is defined in equation (2).

To apply EMC, we must specify the number of chains that are to run in parallel and an associated temperature ladder. These settings are chosen to improve mixing and convergence and can be determined in test runs of the algorithm. We run two independent chains using randomly chosen starting points and examine trace plots of the fitness function. For each chain we use the Raftery–Lewis diagnostic to determine length of burn-in (Raftery and Lewis, 1992). We determine the total number of iterations for the EMC algorithm by the Gelman–Rubin convergence diagnostic (Gelman and Rubin, 1992). These diagnostics are applied to the values of the penalized likelihood (fitness function) associated with the models sampled from the stationary distribution (the chain

with temperature value equal to 1). We also monitor the overall acceptance rates of the EMC sampling algorithm. More details of the EMC algorithm are given in Appendix B.

2.5 Screening of SNPs

Efficiency of stochastic algorithms often diminishes as the total number of models increases. For this reason, we have found it useful to reduce the number of SNPs included in the analysis using a screen when p is larger than about 75. Such a screen will typically be fairly permissive, leaving only the weakest candidates out of the stochastic search. The screen should be quick to calculate, adjust for the same variables and consider the same genetic parametrizations as in the full analysis. In our analyzes, we calculated marginal (i.e. SNP-at-a-time) Bayes Factors for each of the log-additive, dominant and recessive models of association against the model of no association. We ordered SNPs according to the maximum of the three marginal Bayes factors and retained those with a maximum greater than or equal to one. The remaining are SNPs whose data are more likely under one of the genetic models of association than under a model of no association. More detail is available in Appendix C.

2.6 Posterior Inference

We utilize Bayes factors for quantifying association at multiple levels and assessing the most likely SNP-specific genetic parametrization, given the models visited in the EMC sampling algorithm. While posterior probabilities provide a measure of evidence for hypotheses or models, it is often difficult to judge them in isolation as individual model probabilities may be "diluted" as the space of models grows (Clyde and George, 2004). Bayes factors compare the posterior odds of any two models (or hypotheses) to their prior odds and measures the relative strength of evidence in the data for one model, $\mathfrak{M}_{\boldsymbol{\gamma}_1}$, to another, $\mathfrak{M}_{\boldsymbol{\gamma}_2}$, defined as

$$\mathrm{BF}(\mathfrak{M}_{\boldsymbol{\gamma}_1} : \mathfrak{M}_{\boldsymbol{\gamma}_2}) = \frac{p(\mathfrak{M}_{\boldsymbol{\gamma}_1} \mid D)/p(\mathfrak{M}_{\boldsymbol{\gamma}_2} \mid D)}{p(\mathfrak{M}_{\boldsymbol{\gamma}_1})/p(\mathfrak{M}_{\boldsymbol{\gamma}_2})}.$$

See Goodman (1999) for a discussion on the usefulness of Bayes factors in the medical context and Wakefield (2007) for their use in controlling false discoveries in genetic epidemiology studies. Below we define Bayes factors for measuring overall or global significance, gene significance and SNP significance.

2.6.1 Global Bayes Factor

The Bayes factor in favor of H_A, the alternative hypothesis that there is at least one SNP associated with disease, to H_0, the null hypothesis that there is no association between the SNPs under consideration and disease, measures the relative weight of evidence of H_A to H_0. The null hypothesis corresponds to the null model, the model including no SNPs, and is denoted $\mathfrak{M}_0$. The alternative hypothesis is represented by all of the non-null models in $\mathfrak{M}$. Because the space of models is so large, the null model (or any single model in general) may receive small probability, even when it is the highest probability model (the dilution effect of large model spaces), thus Bayes factors are more useful as a measure of evidence.

The Global Bayes factor for comparing H_A to H_0 may be simplified to

$$\mathrm{BF}(H_A : H_0) = \sum_{\mathfrak{M}_\gamma \in \mathfrak{M}} \mathrm{BF}(\mathfrak{M}_{\boldsymbol{\gamma}} : \mathfrak{M}_0) p(\mathfrak{M}_{\boldsymbol{\gamma}} \mid H_A) \tag{3}$$

which is the weighted average of the individual Bayes factors $\mathrm{BF}(\mathfrak{M}_{\boldsymbol{\gamma}} : \mathfrak{M}_0)$ for comparing each model in H_A to the null model with weights given by the prior probability of $\mathfrak{M}_{\boldsymbol{\gamma}}$ conditional on being in H_A, $p(\mathfrak{M}_{\boldsymbol{\gamma}} \mid H_A)$. For a large number of SNPs, it is impossible to enumerate the space of models and posterior summaries are often based on models sampled from the posterior distribution. In equation (3), if we replace the average over all models in H_A with the average over the unique models sampled via the EMC algorithm, $\mathfrak{S}$, the result

$$\mathrm{BF}(H_A : H_0) > \mathrm{BF}_S(H_A : H_0) \equiv \sum_{\mathfrak{M}_\gamma \in \mathfrak{S}} \mathrm{BF}(\mathfrak{M}_{\boldsymbol{\gamma}} : \mathfrak{M}_0) p(\mathfrak{M}_{\boldsymbol{\gamma}} \mid H_A)$$

is a lower bound for the Bayes factor for testing global association. If the lower bound indicates evidence of an association, then we can be confident that this evidence will only increase as we explore more models.

2.6.2 Multilevel Bayes Factors

While it is of interest to quantify association at the global level, interest is primarily in identifying the gene(s) and variant(s) within those genes that drive the association. For this purpose we begin by defining SNP inclusion probabilities and associated Bayes factors. These marginal summaries are adjusted for the other potentially important SNPs and confounding variables and provide a measure of the strength of association at the level of individual SNPs. Given each sampled model

$\mathcal{M}_{\boldsymbol{\gamma}} \in \mathfrak{S}$ and the model specification vectors $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_S)$ previously defined in Section 2.1, the inclusion probability for SNP s is estimated as:

$$p(\gamma_s \neq 0 \mid D) = \sum_{\mathcal{M}_{\boldsymbol{\gamma}} \in \mathfrak{S}} 1_{(\gamma_s \neq 0)}\ p(\mathcal{M}_{\boldsymbol{\gamma}} \mid D, \mathfrak{S}) \tag{4}$$

where $p(\mathcal{M}_{\boldsymbol{\gamma}} \mid D, \mathfrak{S})$ is the posterior probability of a model re-normalized over the sampled model space. The SNP Bayes factor is the ratio of the posterior odds of the SNP being associated to the prior odds of the same, and is defined as:

$$\text{BF}(\gamma_s \neq 0 : \gamma_s = 0) = \frac{p(\gamma_s \neq 0 \mid D)}{p(\gamma_s = 0 \mid D)} \cdot \frac{p(\gamma_s = 0)}{p(\gamma_s \neq 0)};$$

where $p(\gamma_s \neq 0)$ is the prior probability of SNP s being associated. Estimates of the SNP Bayes Factor may be obtained using the estimated SNP inclusion probabilities from (4).

We define the probability of the SNP-specific genetic parametrization conditional upon a SNP being associated as:

$$\begin{aligned}
p(\text{Log Additive} \mid \gamma_s \neq 0) &= \sum_{\mathcal{M}_{\boldsymbol{\gamma}} \in \mathfrak{S}} 1_{(\gamma_s = 1)}\ p(\mathcal{M}_{\boldsymbol{\gamma}} \mid D, \mathfrak{S}), \\
p(\text{Dominant} \mid \gamma_s \neq 0) &= \sum_{\mathcal{M}_{\boldsymbol{\gamma}} \in \mathfrak{S}} 1_{(\gamma_s = 2)}\ p(\mathcal{M}_{\boldsymbol{\gamma}} \mid D, \mathfrak{S}), \\
p(\text{Recessive} \mid \gamma_s \neq 0) &= \sum_{\mathcal{M}_{\boldsymbol{\gamma}} \in \mathfrak{S}} 1_{(\gamma_s = 3)}\ p(\mathcal{M}_{\boldsymbol{\gamma}} \mid D, \mathfrak{S}).
\end{aligned}$$

These probabilities may be used to determine the most likely genetic parametrization.

In cases where there are SNPs in Linkage Disequilibrium (LD), SNP inclusion probabilities may underestimate the significance of an association at that locus. This occurs because SNPs in LD provide competing explanations for the association, thereby diluting or distributing the probability over several markers. Since the amount of correlation between markers across different genes is typically negligible, calculating inclusion probabilities and Bayes factors at the gene level will not be as sensitive to this dilution. We define a gene to be associated if one or more of the SNPs within the given gene are associated. We define the gene inclusion probability as:

$$p(\Gamma_g = 1 \mid D) = \sum_{\mathcal{M}_{\boldsymbol{\gamma}} \in \mathfrak{S}} 1_{(\Gamma_g = 1)}\ p(\mathcal{M}_{\boldsymbol{\gamma}} \mid D, \mathfrak{S});$$

where $\Gamma_g = 1$ if at least one SNP in gene g is in model $\mathbf{\mathcal{M}_\gamma}$ and is zero otherwise. The gene Bayes factor is defined as:

$$\text{BF}(\Gamma_g = 1 : \Gamma_g = 0) = \frac{p(\Gamma_g = 1 \mid D)}{p(\Gamma_g = 0 \mid D)} \cdot \frac{p(\Gamma_g = 0)}{p(\Gamma_g = 1)};$$

where $p(\Gamma_g = 1)$ is the prior probability of one or more SNPs in gene g being associated.

Given the multilevel posterior calculations at the global, gene and SNP level, we are able to determine if a set of genetic markers either within a pathway or a gene of interest, warrants further investigation. In order to carry out inference using posterior inclusion probabilities or Bayes factors, we must specify a prior distribution, $p(\mathbf{\mathcal{M}_\gamma})$, on the space of models $\mathbf{\mathcal{M}}$. While it is common practice to adopt a uniform distribution over the space of models, this choice has the potentially undesirable implication that $\frac{3}{4}$ of the SNPs are expected to be included *a priori*. While a uniform prior distribution leads to 3:1 prior odds that a SNP is included, the prior odds of at least one SNP being included (which is used in the global Bayes factor) are $1 - \frac{1}{4^p}$. It is clear that one cannot have equal prior odds for the different hypotheses of association at the three levels. In the next section we discuss prior distributions over models that are chosen to have desirable operating characteristics at each level of the hierarchy.

2.7 Choice of Prior Distribution on Models

A simple alternative to the uniform prior distribution is to model the inclusion of SNPs as independent Bernoulli variables with a specified probability ρ. This implies *a priori* that the number of included SNPs, $s_{\boldsymbol{\gamma}}$, has a binomial distribution, $\mathsf{Bin}(S, \rho)$ distribution. For large S this results in a fairly concentrated prior distribution on model size around $S\rho$. Rather than taking ρ fixed, we may assign ρ a prior distribution, which provides over-dispersion and added robustness to prior misspecification (Ley and Steel, 2007; Scott and Berger, 2008) Our hierarchical prior distribution on models $\mathbf{\mathcal{M}_\gamma}$ is specified as follows: the prior distribution for ρ is a $\mathsf{Beta}(a, b)$. Conditional upon ρ, the prior distribution on the model size $s_{\boldsymbol{\gamma}}$ is $\mathsf{Bin}(S, \rho)$, which leads to a beta-binomial distribution on the model size. Finally, conditional upon $s_{\boldsymbol{\gamma}}$, the number of SNPs included in $\mathbf{\mathcal{M}_\gamma}$, and ρ, the prior distribution on the genetic parametrization of the included SNPs is a uniform distribution across all possible genetic parametrization.

2.8 Prior Hyper-Parameter Choice

The choice of the beta-binomial hyper-parameters a, b influence the prior expected model size (and variance), the penalty in the penalized likelihood used in the sampling algorithm and computation of posterior probabilities, and the prior odds in the computation of the global, gene and SNP level Bayes factors. As we use Bayes factors in making a decision of association, the choice of hyper-parameters affects false and true positive rates. In what follows, we estimate the false/true positive rates as a function of a and b via simulation, using these results to guide choice of hyper-parameters for analysis of the experimental data.

2.8.1 Characteristics of Prior Distributions Given Hyper-parameters

Figure 1 illustrates plots of the prior mass assigned to all models of size s in the left hand column and the prior probability of an individual model given that its size is $s_{\boldsymbol{\gamma}}$ in the middle column. For all calculations we assume that $S = 66$, roughly the number of SNPs that passed the screen in the analysis of the experimental data presented in Section 3. From top to bottom, the rows of Figure 1 correspond to priors $b = 1$, $b = S/2$ and $b = S$, respectively. The color of the curve represents values of a in $(1, 1/2, 1/4, 1/8, 1/16, 1/32)$. When $a = 1$ and $b = 1$, the prior distribution on model size is uniform. Although this "non-informative" prior distribution may seem to be desirable, the implied prior distribution on the individual models is not monotone decreasing in $s_{\boldsymbol{\gamma}}$, but has a mode at 0 and at S. As b increases and a decreases, the prior distribution on individual models becomes more monotonic. Prior distributions with $b > 1$ place more mass on smaller models.

The right hand column of Figure 1 plots the relative penalty – the penalty of any model of size S divided by the penalty of the null model – against S as a function of the hyper-parameters. These plots also include the relative penalties induced by using AIC and BIC without the added penalty from the prior distribution on the model space. For all values of a and b plotted, the relative penalties incorporating the model prior are similar to BIC for a model size of approximately 20 or 30 (or less). However, the beta-binomial prior penalizes larger models less than BIC. When $b = S$, the relative penalty induced by the beta-binomial prior distribution is comparable to BIC across the full range of model sizes.

The prior distribution over the model space enters the calculation of the global Bayes factor in two places: it appears in the relative penalized likelihood described above and in the global prior

odds of the null to the alternative hypotheses. Under the beta-binomial prior distribution, the global prior odds of the null to the alternative hypothesis are:

$$\frac{p(H_0)}{p(H_A)} = \frac{\Gamma(a+b)\Gamma(b+S)}{\Gamma(b)\Gamma(a+b+S) - \Gamma(a+b)\Gamma(b+S)}.$$

Figure 2 shows the prior odds and the expected model size as a function of a and b. The prior odds increase and the model size decreases rapidly as b increases and a decreases.

Similarly, the SNP specific Bayes factor depends on the model prior probability through the relative penalized likelihood and through the prior odds. Under the the beta-binomial prior distribution, the prior odds of SNP s being associated are

$$\frac{p(\gamma_s > 0)}{p(\gamma_s = 0)} = \frac{a}{b}$$

thus, as we increase b or decrease a, the prior odds of any given SNP being associated will decrease.

2.8.2 Selection of Hyper-parameters via Simulation

In this section, we describe a simulation–based procedure for choosing the hyper–parameters a and b best suited to a particular analysis. The procedure requires *a priori* specification of the expected number and magnitude of associations among the variants under study. We generate a set of 30 simulated data sets and use these to estimate global, gene– and SNP–specific Bayes factors and use the ensemble of data sets to estimate true and false discovery rates as a function of a and b.

The simulated data sets are structured so as to reflect the details — genes, tag SNPs, LD structure, and sample size — of a candidate pathway in the North Carolina Ovarian Cancer Study (NCOCS) (J.M. Schildkraut (2008)), a population based case-control study. We delay presenting additional details of the study until Section 3. The pathway of interest is comprised of 53 genes tagged by 508 tag SNPs.

We simulate the genotypes in two stages. First, for each of the 53 genes represented in the data set, we phase the NCOCS control SNP genotype data and estimated recombination rates using PHASE (Stephens et al., 2001). Second, given a model of association, we generate case–control data at these tags using HAPGEN (Marchini and Su, 2006). Ten of the simulations are null; there are no associations in the genes of interest. The remaining twenty simulations assume that a randomly chosen subset of 9 genes are associated and that within the associated genes, a single, randomly chosen tag is the source of the association. Within ten of these associated simulations,

three of the associated tag SNPs are accorded an odds ratio (OR) of 1.25, three an OR of 1.5 and three an OR of 1.75 and within the other ten associated simulations, three of the associated tag SNPs are accorded an OR of 1.75, three with an OR of 2.0 and three with an OR of 2.25. In each of the twenty associated simulations, one SNP with each OR is assumed to have a dominant genetic parametrization, one of each a log–additive, and one of each a recessive.

Based on convergence diagnostics, we ran the EMC algorithm for 400,000 iterations. We set $N = 5$ parallel chains over a temperature ladder spanning a minimum temperature of 0 and a maximum temperature of 10. We reduced the number of SNPs in each simulation to a subset with a marginal Bayes factor in favor of association estimated to be 1.0 or above as described in Appendix C, leaving us with a range of 37 to 74 SNPs per simulation.

We ran the EMC sampling algorithm using a set of (a, b) to span a range of prior odds as discussed in Section 2.8.1; we chose values for $b \in (S/2, S, 3S/2)$ and values for $a \in (1, 1/4, 1/8, 1/32)$. Given these values for the hyper-parameters, true and false positive rates

$$\mathsf{TP} = \frac{\#\ \text{associated SNPs declared associated}}{\#\ \text{associated SNPs}} \tag{5}$$

$$\mathsf{FP} = \frac{\#\ \text{unassociated SNPs declared associated}}{\#\ \text{unassociated SNPS}} \tag{6}$$

were estimated based on pooling decisions over all 30 simulations. SNPs were declared associated if the SNP specific Bayes factor $\mathrm{BF}(\gamma_s > 0 : \gamma_s = 0)$ exceeded a threshold of 10, corresponding to strong evidence of an association based on Jeffrey's scale (Jeffreys, 1961, page 432) as modified by Kass and Raftery (1995); we include these grades of evidence in Table 2. The true and false positive rates are calculated for the global Bayes factors $\mathrm{BF}(H_a : H_0)$ and gene Bayes factors, $\mathrm{BF}(\Gamma_g = 1 : \Gamma_g = 0)$, substituting pathway and gene for SNP in the above definitions. We used a threshold of 10 for the pathway (global) and gene-specific Bayes factors. We calculate these rates using the SNPs that passed the initial marginal Bayes factor screen and with all SNPs. We compare the joint multi-level inference to the marginal Bayes factor model. In order to facilitate this comparison, for the marginal analysis, we define a SNP to be associated if the maximum marginal Bayes factor over the three genetic models exceeds 10; we declare a gene to be associated if any SNP within the gene is declared associated; and we declare the pathway (global) to be associated if any genes are declared associated.

3 Results

In this section we describe an analysis analyze polymorphic variants of a candidate pathway in the ongoing NCOCS ovarian cancer case–control association study. The NCOCS covers a 48 county region of North Carolina. Cases are between 20 and 74 years of age and were diagnosed with primary invasive or borderline epithelial ovarian cancer after January 1, 1999. Controls are frequency matched to the cases by age and race and have no previous diagnosis of ovarian cancer. In the analysis we present, we focus on self–reported Caucasians and a histological subtype of the cancers, giving us a total of 397 cases and 787 controls. Because the ovarian cancer results have not yet been published, we anonomyze the pathway, the genes chosen to represent it and the IDs of the SNPs tagging variation in those genes. The pathway is comprised of 53 genes tagged by 508 tag SNPs.

3.1 Prior Choice and Model Characterization via Simulation

Before proceeding with the analysis of the experimental data, we must first specify the hyperparameters a and b of the beta-binomial prior distribution on models. Using the simulation methods described in Section 2.8.2, we select hyper-parameters to control the proportion of false positives and true positives at the global, gene and SNP level. Table 1 shows the true and false positive rates as a function of a and b. Our choice for the hyper-parameters, $a = 1/8$ and $b = S$, is highlighted in red in Table 1. The hyper-parameters were chosen to strike a balance between the proportion of false positives and true positives on the global, gene and SNP level. From the simulation results, this prior is characterized by a high true positive rate, detecting 91.5% of the associated genes and 83.9% of the associated SNPs that passed the screen. We were willing to accept a higher false positive rate (28.3% at the gene level and 23.0% at the SNP level) as detected SNPs would be followed up for independent validation. While there is a limit to the number of SNPs that can be validated, false negatives are lost to any planned follow up and are thereby viewed as being more costly. Out of all null simulations, only 10% were falsely declared to have at least one associated SNP, while all associated simulations were detected.

The false and true positive rates reported above pertain to SNPs that passed the marginal screen, and do not account for errors made by the screen. Overall, the marginal Bayes factor screen has 1.0 % and 3.8% SNP and gene false positive rates, respectively, with 43.3% (SNP) and 47.8%

(gene) true positive rates. Taking into account the screen, MISA has slightly higher false positive rates 2.1% and 11.7% at the SNP and gene levels, but higher true positive rates 54.4% (SNP) and 70.0% (gene). The screen eliminates many clearly true negatives thereby keeping the overall false positive rate low. While the marginal screen could be used for the entire analysis, MISA is able to detect an additional 25% true associations at the SNP level and an additional 46% at the gene level.

Figure 3 depicts boxplots of the gene and SNP Bayes factors as a function of the ORs used in the simulation study. The dotted line in the figures corresponds to the Bayes factor threshold (BF > 10 or BF $< 1/10$) used to declare a strong association using the modified Jeffreys' scale of evidence as seen in Table 2. Given the sample size and characteristics of the NCOCS study, we have power of about 75% (or higher) to detect odds ratios of 1.5 and higher; on the other hand there is about 50% power to detect an association with ORs of 1.25. While using a lower Bayes factor threshold to detect SNPs would increase the true positive rate and power, it would also increase the number of false positives.

The prior expected model size given this choice of a and b and for S ranging from 37 to 74 (the range of the number of SNPs included in our simulation studies) is approximately 0.125, the prior odds of the null model to the alternative model is approximately 11 to 1 and the prior probability of including any SNP in any model is approximately 0.002. This value takes into account the correlation structure among SNPs, our costs/benefits of false/true positives, and reflects the fact that this is a candidate gene study as opposed to a genome-wide association study with a much larger S.

3.2 Ovarian Cancer Analysis

Using the hyperparameters $a = 1/8$ and $b = S$, we used MISA to identify associated SNPs and genes in the candidate pathway for the NCOCS case-control data. All models included the patients age and an indicator variable for previous diagnosis of breast cancer as potential confounding variables. Model search was confined to the identifying SNPs and their genetic model parametrization. We used the same parameters for the model fitting algorithm as in the simulation study (400,000 iterations, a population size of $N = 5$ and a temperature ladder from 0 to 10). As in the simulation, we screened 508 SNPs using marginal Bayes factors, yielding 70 SNPs that exceeded the threshold

of 1 in favor of an association.

We ran two independent runs of the algorithm from independent starting points for a total of $400,000$ iterations for each run and convergence was determined as in the simulation study by assessing the Gelman-Rubin shrink factor (Gelman and Rubin, 1992) over the iterations and by investigating the convergence of the values of the global and marginal Bayes factors across the iterations and between the two independent runs of the algorithm. While these diagnostics suggested the separate chains had reached convergence (Gelman-Rubin shrink factor approximately one), a plot of the marginal inclusion probabilities from two independent runs suggested that these quantities had not yet converged. We ran the two chains for an additional 1.2 million iterations and reassessed convergence (Figure 7).

We estimate the pathway–wide Bayes factor for association to be $\mathrm{BF}(H_A : H_0) = 4.30$ which constitutes "substantial" evidence in favor of an association between the pathway and ovarian cancer based on Jeffreys grades of evidence that can be referenced in Table 2. To put this value into perspective, Jeffreys states that for values of the Bayes factor greater than 10, we can have "strong confidence that the results will survive future investigations". We should note that our estimate of the global Bayes factor is a lower bound of the Bayes factor that would be obtained by enumeration of all models.

Figures 4 and 5 summarize the associations of the 20 most highly associated of the 70 SNPs and the genes they represent. SNPs and genes in the pathway are denoted by a two level name (eg. S1 and G1) where the number represents the rank of the SNP or gene by its respective Bayes factor. These plots summarize the top 100 models $\mathcal{M}_{\boldsymbol{\gamma}} \in \mathcal{M}$ selected on basis of their posterior model probabilities. Models are ordered on the x–axis in descending probability and the width of the column associated with a model is proportional to that probability. SNPs (Figure 4) or genes (Figure 5) are represented on the y–axis. The presence of a SNP or gene in a model is indicated by a colored block at the intersection of the model's column and the SNP's or gene's row. In Figure 4, the color of the block indicates the parametrization of the SNP: purple for log–additive, blue for recessive and red for dominant. SNPs one to six have Bayes factors greater than 10, providing strong evidence for an association between these SNPs and ovarian cancer.

The top 10 models depicted in Figure 4 include only a single SNP in addition to the design variables that are incorporated in all models (the design effects are not pictured in the figure). The

top model includes only the log–additive genetic parametrization of SNP S1 in gene G1 . Under this model we estimate the OR for S1 to be approximately 1.42 (the posterior mode). The second ranked sampled model is comprised only of the log–additive genetic parametrization of SNP S2 in gene G1 and the design variables, with an estimated OR for S2 of 1.37. The magnitude of the effect sizes of both S1 and S2 are relatively small relative to the power of the study and are on the lower end of the range of ORs considered in the simulation study.

Figure 4 also illustrates that many of the top models beyond the first ten include multiple SNPs. This suggests that if we were to restrict our attention to single SNP models (as is done with many analytical procedures in use) we would potentially lose substantial information regarding their joint effects. For example, model 11 is comprised of both SNP S3 from gene G4 and SNP S1 from gene G1, while model 14 is comprised of both SNP S3 from gene G4 and SNP S2 from gene G1. In both cases, SNP S3 is included in models with a SNP from gene G1. This may indicate that not only are SNPs S1, S2, and S3 important as single effects in the top 3 models, but that their association with ovarian cancer may be in some ways interrelated and that their combined effects may be of interest. Further, in some cases multiple variants within a gene may better tag the underlying variant of interest.

Taking into account model uncertainty by averaging over inclusion of other potentially associated SNPs and their genetic parametrizations, the SNP Bayes factors of S1 (BF = 49.6) and S2 (BF = 37.1) provide very strong evidence that the SNPS are associated with ovarian cancer, while the Bayes factor for S3 (BF = 20.8) provides strong evidence for an association. Based on Jeffreys' grades of evidence, these results are likely to stand up to further study.

Both SNP 1 and SNP 2 are in gene G1 which has a gene Bayes factor of 31.65, giving very strong evidence that at least one of the SNPs in gene 1 is associated with ovarian cancer. Figure 6 provides a detailed look at the SNPs within gene G1, illustrating inclusion of the 3 G1 SNPs and of gene G1 as a whole in the top 100 sampled models (top panel) as well as summaries related to the joint inclusion of SNPs (bottom panel). Note that when one of SNP S1 or S2 is included in a model, the other is often not (at least in the top 50 models). This trade off often arises when SNPs are highly correlated (i.e. in high linkage disequilibrium). In Figure 6, the upper triangle of the bottom plot depicts LD, here measured as squared correlation. SNPs S1 and S2 have a modest LD of 0.5 whereas the correlation between SNPs S1 and S57 and between SNPs S2 and S57 is

approximately 0.1. The lower triangle of this plot depicts the estimated joint Bayes factor defined as the ratio of the posterior odds that both of the indicated pair of SNPs is included to the prior odds that both SNPs are included. While the upper plot of SNP inclusion suggests that models with both S1 and S2 are less likely than models with only one of the pair, the joint Bayes factor in favor of both S1 and S2 being included implies that the posterior odds for their inclusion greatly exceeds their prior odds.

Gene G1 highlights a shortcoming of the use of Bayes factors with nested hypotheses. Note that the gene Bayes factor for G1 (Figure 6) is smaller than each of the SNP Bayes factors for S1 and S2. The posterior probability that gene G1 is associated is based on adding the probabilities of all models that include at least one SNP from that gene, hence the *posterior probability* for gene inclusion is always greater than or equal to the probability that any one SNP is included (i.e. posterior probabilities observe a monotonicity property with nested hypotheses). Bayes factors, however, are the ratio of posterior odds to prior odds, and do not share this property. Because the prior distribution controls for multiplicities of testing with a large number of SNPs, the prior probabilities of SNP and gene inclusion and the resulting posterior probabilities may be very small. Thus while posterior probabilities enjoy the monotonicity property, judging their importance in absolute terms is difficult as a result of the dilution of probabilities when there are many models under consideration. While not monotonic, the Bayes factor allows one to put the evidence of association in context by contrasting the posterior to the prior odds of inclusion. Despite the lack of Bayes factor monotonicity, the conclusion remains that evidence is very strong that gene G1 is associated with ovarian cancer.

4 Discussion

In this paper, we describe an analytic strategy for hypothesis based association studies that allows one to quantify evidence of association at 3 levels: global (e.g. pathway–wide), gene, and SNP while allowing for uncertainty in the genetic parametrization of the markers. Our methodology improves upon marginal, snp–at–a–time methods by avoiding their implicit conditional independence assumption, by considering multivariate adjusted associations, by providing a natural framework for multi–level inference and by providing an explicit, tunable and characterizable multiple comparisons correction. Indeed, we show through a simulation that MISA used after a marginal screen

has better operating characteristics than the marginal analysis alone.

Like logic regression (Kooperberg and Ruczinski, 2004), MISA involves a stochastic model search over a space of prospective models for the outcome of interest; but, unlike logic regression, it focuses entirely on identifying main effects and specific genetic models for association and includes explicit procedures for calculating important hypothesis testing quantities. These include Bayes factors for association at the SNP, gene and pathway levels. In addition to being of interest in their own right for their ability to address basic research hypotheses, the hierarchy of Bayes factors can be used to elucidate the nature of associations. This is especially true at the level of genes and SNPs where the significance of individual SNPs, as measured by their inclusion probabilities and Bayes factors, may be diluted due to their correlation with neighboring SNPs in that gene. Gene level inference implicitly accounts for these correlations and is robust to this effect. SNPs S1 and S2 within gene G1 of the ovarian cancer study are an example.

Our methodology can be readily adapted to accommodate the analysis of specific studies through choice of hyper–parameters, different study designs and outcomes. The use of study based–simulations plays a critical role in our methodology by allowing the investigator to choose the model's hyper–parameters on basis of their expected operating characteristics in the study of interest. MISA can also be adapted to accommodate models for other forms of outcome variables. For example, likelihoods such as used to model quantitative traits or survival data may be used in place of the logistic regression model. Missing SNP genotype data is a common phenomenon; we address this problem by generating multiple imputations for the missing data using existing genetic analysis software. This allows us to use all study subjects while reporting valid uncertainty summaries.

Whether the study at hand is one that will be followed up with a pointed evaluation of a small number of candidates or with a more complete second phase study, MISA's multilevel Bayes factors can be used to determine the pathways, genes, regions within genes or individual variants that would benefit most from further investigation.

Acknowledgments

This work was supported by the Duke SPORE in Breast Cancer, P50-CA068438; the North Carolina Ovarian Cancer Study, R01-CA076016; NIH/NHLBI R01-HL090559; NSF DMS-0342172; and

NSF DMS-0422400. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendices

A Implied Prior Distribution under AIC

Given that a closed form expression for the marginal likelihood is not available for logistic regression, we have used AIC to approximate the likelihood. In what follows, we determine a prior distribution on model coefficients that is consistent with AIC.

We assume a normal prior distributions on the d_γ dimensional vector of regression coefficients (log odds ratios) of the form

$$p(\boldsymbol{\theta_\gamma}|\mathcal{M}_{\boldsymbol{\gamma}}) \sim \mathsf{N}\left(\boldsymbol{t_\gamma}, \frac{1}{k}\mathcal{I}_{\boldsymbol{\gamma}}^{-1}\right),$$

where $\mathcal{I}_{\boldsymbol{\gamma}}$ is the observed Fisher information under model $\mathcal{M}_{\boldsymbol{\gamma}}$ evaluated at the maximum likelihood estimates (MLEs) $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$. Setting the covariance matrix to be proportional to the inverse Fisher information ensures that the correlation structure in the prior distribution matches that of the likelihood.

In order to approximate the marginal likelihood we use a Laplace approximation based on expanding the log likelihood in a second order Taylor's series expansion about $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$:

$$\mathcal{L}(\boldsymbol{\theta_\gamma} \mid \mathcal{M}_{\boldsymbol{\gamma}}) \approx \mathcal{L}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}} \mid \mathcal{M}_{\boldsymbol{\gamma}}) - \frac{1}{2}(\boldsymbol{\theta_\gamma} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}})^T \mathcal{I}_{\boldsymbol{\gamma}}(\boldsymbol{\theta_\gamma} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}})$$

leading to the approximate marginal likelihood

$$\begin{aligned} p(D \mid \mathcal{M}_{\boldsymbol{\gamma}}) &\approx \exp\{\mathcal{L}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}} \mid \mathcal{M}_{\boldsymbol{\gamma}})\} \int \exp\{-\frac{1}{2}(\boldsymbol{\theta_\gamma} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}})^T \mathcal{I}_{\boldsymbol{\gamma}}(\boldsymbol{\theta_\gamma} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}})\} \times \\ &\qquad \frac{1}{(2\pi)^{\frac{d_\gamma}{2}}} |k\mathcal{I}_{\boldsymbol{\gamma}}|^{\frac{1}{2}} \exp\{-\frac{k}{2}(\boldsymbol{\theta_\gamma} - \boldsymbol{t_\gamma})^T \mathcal{I}_{\boldsymbol{\gamma}}(\boldsymbol{\theta_\gamma} - \boldsymbol{t_\gamma})\} d\boldsymbol{\theta_\gamma} \\ &= \left(\frac{k}{k+1}\right)^{\frac{d_\gamma}{2}} \exp\{\mathcal{L}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}})\} \exp\{-\frac{1}{2}\frac{k}{k+1}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}} - \boldsymbol{t_\gamma})^T \mathcal{I}_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}} - \boldsymbol{t_\gamma})\}. \end{aligned}$$

Setting the approximate $\log(p(D \mid \mathcal{M}_{\boldsymbol{\gamma}})$ equal to -0.5AIC we have equality when the prior mean $\boldsymbol{t_\gamma}$ is set to $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$ causing the right most term to vanish and $k = \frac{1}{\exp(2)-1}$. Roughly speaking, this implies

that the prior standard deviation of any standardized log odds ratio is about 2.5. Even if the prior mean were zero, this provides enough dispersion to cover the range of log odds ratios anticipated in practice.

B Evolutionary Monte Carlo

We use an Evolutionary Monte Carlo (EMC) (Liang and Wong, 2000) algorithm to sample models that maximize a fitness function $\psi(\mathcal{M}_{\boldsymbol{\gamma}})$. In the current setting, this quantity is proportional to the log of the posterior probability of the sampled models. This is achieved using parallel tempering with N parallel Markov chains, each associated with a decreasing temperature T_i in a temperature ladder $T = \{T_1, T_2, \cdots, T_N\}$ and sampling from the distribution $p_{T_i}(\mathcal{M}_{\boldsymbol{\gamma}} \mid D) \propto \exp(\frac{\psi(\mathcal{M}_{\boldsymbol{\gamma}})}{T_i})$. The advantages of parallel tempering over single chain MCMC methods include improved mixing and its ability to escape local modes. The resulting sample from the model space $\mathfrak{M}$ (the sample of models from the chain with temperature $T_i = 1$) is from the stationary (posterior) distribution. In the EMC framework, each current state of one of the parallel chains corresponds to an "individual" or model and the full set of current states of the chains correspond to the "population" or set of models.

The parallel chains are updated by the following populations moves that are based on a genetic algorithm: Mutation, Crossover and Exchange. For each update we accept or reject the proposed move on the population of models based on the probability $\min(1, r)$. Here, r is the Metropolis–Hastings (MH) ratio corresponding to the original and updated populations, $P_{\mathfrak{M}}$ and $P_{\mathfrak{M}*}$ respectively, that has the general form:

$$r = \frac{f(P_{\mathfrak{M}*})}{f(P_{\mathfrak{M}})} \frac{t(P_{\mathfrak{M}} \mid P_{\mathfrak{M}*})}{t(P_{\mathfrak{M}*} \mid P_{\mathfrak{M}})},$$

where $f(P_{\mathfrak{M}})$ is the product joint distribution for the population of models defined as

$$f(P_{\mathfrak{M}}) = \prod_{i=1}^{N} p_{t_i}(\mathcal{M}_i \mid D),$$

and $t(P_{\mathfrak{M}} \mid P_{\mathfrak{M}*})$ is the transition probability between populations.

We first update the population via a mutation step where we perform a Metropolis update on the population by choosing a model, or current value of one of the chains and taking one of the SNP indicators and mutating its status in the chosen model. Given our population of models,

$P_{\mathfrak{M}} = \{\mathfrak{M}_1, \mathfrak{M}_2, \cdots, \mathfrak{M}_p\}$, we sample one of the models $\mathfrak{M}_{\boldsymbol{\gamma}}$, and mutate it to some model $\mathfrak{M}_{\boldsymbol{\gamma}}*$ and accept or reject the new population, $P_{\mathfrak{M}}* = \{\mathfrak{M}_1, \cdots, \mathfrak{M}_{\boldsymbol{\gamma}}*, \cdots, \mathfrak{M}_p\}$ based on the probability $\min(1, r_m)$, where r_m is the MH ratio for the mutation update. Specifically, to update the model $\mathfrak{M}_{\boldsymbol{\gamma}}$ to model $\mathfrak{M}_{\boldsymbol{\gamma}}*$ we update the corresponding model specification vector $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_p)$ to $\gamma*$ by mutating one value $\gamma_m \in \{0, 1, 2, 3\}$, based on the corresponding SNP inclusion and genetic mode of inheritance in the model, to another value $\gamma_m* \in \{0, 1, 2, 3\}/\gamma_m$. We then choose γ_m* based on the following probabilities for each possible value of γ_m*:

$$p(\mathfrak{M}_{\boldsymbol{\gamma}}* \mid \gamma_m*) = \frac{\exp\left(\psi(\mathfrak{M}_{\boldsymbol{\gamma}}* \mid \gamma_m*)/T_\gamma\right)}{\sum_{\mathfrak{M}_{\gamma}* \neq \mathfrak{M}_{\gamma}} \exp\left(\psi(\mathfrak{M}_{\boldsymbol{\gamma}}* \mid \gamma_m*)/T_\gamma\right)}.$$

We accept or reject the new population $P_{\mathfrak{M}}*$ based on the MH ratio

$$r_m = \frac{\sum_{\mathfrak{M}_{\gamma}* \neq \mathfrak{M}_{\gamma}} \exp\left(\psi(\mathfrak{M}_{\boldsymbol{\gamma}}* \mid \gamma_m*)/T_\gamma\right)}{\sum_{\mathfrak{M}_{\gamma} \neq \mathfrak{M}_{\gamma}*} \exp\left(\psi(\mathfrak{M}_{\boldsymbol{\gamma}}* \mid \gamma_m*)/T_\gamma\right)}.$$

The population is also updated via the normal parallel tempering exchange step that allows models to move up or down the temperature ladder. We choose a temperature ladder of the form $T_j - T_i = \exp(\frac{T_j}{T_i})$ where $T_i = 1$ for some value i in the ladder. Here, given the current population $P_{\mathfrak{M}} = \{\mathfrak{M}_1, \cdots, \mathfrak{M}_i, \cdots, \mathfrak{M}_j, \cdots, \mathfrak{M}_p\}$, we sample two of the models $\mathfrak{M}_i$ and $\mathfrak{M}_j$ and propose a new population $P_{\mathfrak{M}}* = \{\mathfrak{M}_1, \cdots, \mathfrak{M}_j, \cdots, \mathfrak{M}_i, \cdots, \mathfrak{M}_p\}$, by making an exchange between $\mathfrak{M}_i$ and $\mathfrak{M}_j$ without changing the associated temperature ladder. We only allow the exchange to take place between two models with neighboring temperatures and therefore the transition probability is symmetric with $t(P_{\mathfrak{M}} \mid P_{\mathfrak{M}}*) = t(P_{\mathfrak{M}}* \mid P_{\mathfrak{M}})$. We then accept or reject the new population $P_{\mathfrak{M}}*$ based on the MH ratio

$$r_e = \exp\left((\psi(\mathfrak{M}_j) - \psi(\mathfrak{M}_i))(T_i^{-1} - T_j^{-1})\right).$$

The exchange update allows better models to move down the ladder where the chains explore local perturbations to them, while less interesting models move up the ladder to serve as the foundation for more global perturbations.

The mutation and exchange step make up the normal population updates involved in parallel tempering. Evolutionary Monte Carlo introduces a crossover update inspired by genetic algorithms. While the exchange step is a full state swap, the crossover update allows the states to swap partially. The general idea is that one of the top current models is chosen to "mate" with another random model and two new models are formed by some composition of the two parental models. Thus,

given the current population $P_{\mathfrak{M}} = \{\mathfrak{M}_1, \cdots, \mathfrak{M}_i, \cdots, \mathfrak{M}_j, \cdots, \mathfrak{M}_p\}$, we choose two models $\mathfrak{M}_i$ and $\mathfrak{M}_j$ based on a weighted selection procedure and update these models to $\mathfrak{M}_i*$ and $\mathfrak{M}_j*$ based on randomly switching the values of the two model specification vectors. We then accept or reject the new population $P_{\mathfrak{M}}* = \{\mathfrak{M}_1, \cdots, \mathfrak{M}_i*, \cdots, \mathfrak{M}_j*, \cdots, \mathfrak{M}_p\}$ based on the MH ratio

$$r_c = \exp\left((\psi(\mathfrak{M}_i*) - \psi(\mathfrak{M}_i))/T_i + (\psi(\mathfrak{M}_j*) - \psi(\mathfrak{M}_j))/T_j\right).$$

C Marginal Bayes Factor Screen

We used Laplace approximations to estimate the marginal Bayes Factors (BFs) used to screen the SNPs Kass and Raftery (1995). In particular, we estimated the marginal likelihood of each of the three genetic models of association (log-additive, dominant and recessive) and under the null model (model of no genetic association). The BF for a model of association is defined as the ratio of the marginal likelihood (ML) of that model of association to the ML of the null model.

We accounted for missing genetic data by averaging MLs over the 100 imputed genetic data sets. This affected only the ML calculations under the three genetic models of association, but not the null ML. Hence the BF for an association was computed as the average of imputation–specific BFs.

In the ovarian cancer analysis, the model for each SNP was a logistic regression for disease status given the variables age, an indicator for prior diagnosis of breast cancer and the model specific genotype variable. Age and prior diagnosis of breast cancer were included in all models, including the 'null' model of no association. The simulation models were unadjusted as no design or confounder variables were simulated. We placed a normal, mean zero, standard deviation two prior on the parameter of the genetic effect variable and flat, improper priors on the remaining parameters. We ordered SNPs according to the maximum of the three Bayes factors and considered those with a maximum greater than or equal to one in the EMC model search.

D Convergence Diagnostics

We assess convergence of the sampling algorithm by using graphical diagnostics that summarize two independent, long runs of the algorithm. These diagnostics are plotted in Figure 7 for two independent runs of the ovarian cancer candidate pathway, each 1.2 million iterations long. The

upper left panel depicts end–to–end trace plots of the cost values associated with the models sampled in the two runs. The first is plotted in blue and the second in red. This plot is used to verify that the algorithm's movement around the model space is adequate. If this is the case, the sampled model will change frequently and the associated cost values will vary around an average cost so that points in the plot appear to fall within a common horizontal band. If there appears to be drift at the start of either of the trace plots, the associated iterations, representing a period of burn–in, should be removed from the analysis. In addition, if the sampled cost value changes infrequently, indicating that the algorithm is not moving adequately, the algorithm should be restarted with a larger maximum temperature or a larger number of parallel chains so that adjacent chains can communicate better.

The upper right panel in the figure depicts a plot of the Gelman–Rubin shrink factor (Gelman and Rubin, 1992) computed for the sampled cost values. The shrink factor should approach 1 as the cost values of the models converge. The lower left panel of the figure shows the global Bayes factor for each of the two independent runs (the first in blue and the second in red) as a function of iteration. The estimated Bayes factor, which is a lower bound on the value we would compute were we able to enumerate all models, will increase with every new model sampled. As unique models are sampled less frequently as the algorithm runs, the Bayes factor will begin to plateau suggesting that the value of additional samples is diminishing. The plots of the two runs should begin to converge, with neither making large jumps in the later iterations. Finally, the lower left panel plots the inclusion probabilities for each of the SNPs for the two independent runs of the algorithm. This plot enables us to determine if the values of the marginal SNP inclusion probabilities are consistent across two independent runs of the algorithm. Points in the scatter plot will be near the diagonal when MCMC sampling variability for estimating the associated inclusion probabilities is low. If the plot shows significant deviation from the diagonal, the algorithm has not yet converged and should be allowed to run longer.

Web Resources

The URL for the website presented herein is as follows:

`http://stat.duke.edu/gbye/MISA.html`

References

Clyde M, George EI (2004). Model Uncertainty. Statist. Sci. 19:81–94.

Consortium TWTCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Efron B (2007). Correlation and Large-Scale Simultaneous Significance Testing. Journal of American Statistical Association 102:93–103.

Gelman A, Rubin D (1992). Inference from iterative simulation using multiple sequences. Statistical Science 7:457–472.

Geyer C (1991). Markov chain Monte Carlo maximum likelihood. In *Proc. 23rd Symp. Interface*, pp. 156–163. Computing Science and Statistics.

Goodman S (1999). Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. Annal Intern. Med. 130:1005–1013.

Holland J (1975). Adaptation in Natural and Artificial Systems. The U. of Michigan Press.

Jeffreys H (1961). Theory of Probability (Third ed.). Oxford Univ. Press.

J.M. Schildkraut ea (2008). Cyclin E Overexpression in Epithelial Ovarian Cancer Characterizes an Etiologic Subgroup. Cancer Epidemiology Biomarkers and Prevention 17:585–593.

Kass RE, Raftery AE (1995). Bayes factors. J. Amer. Statist. Assoc. 90:773–795.

Kooperberg C, Ruczinski I (2004). Identifying Interacting SNPs Using Monte Carlo Logic Regression. Genetic Epidemiology 28:157–170.

Ley E, Steel M (2007). On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression.

Liang F, Wong W (2000). Evolutionary Monte Carol:Applications to C_p Model Sampling and Change Point Problem. Statistica Sinica 10:317–342.

Marchini J, Su Z (2006). HAPGEN, a C++ program for simulating case and control SNP haplotypes.

Raftery A, Lewis S (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. Statistical Science 7:493–497.

Raftery AE (1986). A note on Bayes factors for log-linear contingency table models with vague prior information. J. Roy.Statist. Soc. Ser. B 48:249–250.

Ruczinski I, Kooperberg C, LeBlanc M (2003). Logic Regression. Journal of Computational and Graphical Statistics 12:475–511.

Schwender H, Ickstadt K (2007). Identification of SNP interactions using logic regression. Biostatistics 9:187–198.

Scott JG, Berger JO (2008). Multiple Testing, Empirical Bayes, and the Variable-Selection Problem. Discussion Paper 2008-10, Duke University Department of Statistical Science.

Stephens M, Smith N, Donnelly P (2001). A New Statistical Method for Haplotype Reconstruction from Population Data. The American Journal of Human Genetics 68:978–989.

Storey J (2002). A Direct Approach to False Discovery Rates. Journal of the Royal Statistical Society 64:479–498.

Tyrer J, Pharoah P, Easton D (2006). The Admixture Maximum Likelihood Test: A Novel Experiment-wise Test of Association Between Disease and Multiple SNPs. Genetic Epidemiology 30:636–643.

Wacholder S (2004). Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies. Journal of the National Cancer Institute 96:434–442.

Wakefield J (2007). A Bayesian Measure of the Probability of False Discovery in Genetic Epidemiology Studies. The American Journal of Human Genetics 81:208–227.

Tables

Table 1: False and true positive rates at the global, gene and SNP levels as a function of different values for the hyper-parameters a and b. False and true positive rates are calculated based on thresholding the Bayes factors at 10. The selected hyperparameters for our study are highlighted in red.

		False Positive			True Positive		
		Global	Gene	SNP	Global	Gene	SNP
a=1	b=.5S	0.3	0.175	0.122	1.0	0.788	0.695
a=1	b=S	0.3	0.152	0.117	1.0	0.780	0.695
a=1	b=1.5S	0.2	0.140	0.114	1.0	0.771	0.703
a=1/4	b=S	0.1	0.207	0.179	1.0	0.856	0.771
a=1/8	b=S	0.1	0.283	0.239	1.0	0.915	0.839
a=1/32	b=S	0.0	0.366	0.309	1.0	0.924	0.881

Table 2: Jefferys grades of evidence (Jeffreys, 1961, page 432)

Grade	BF($H_A : H_0$)	Evidence against H_0
1	1 to 3.2	Not worth more than a bare mention.
2	3.2 to 10	Substantial
3	10 to 31.6	Strong
4	31.6 to 100	Very Strong
5	> 100	Decisive

Figures and Legends

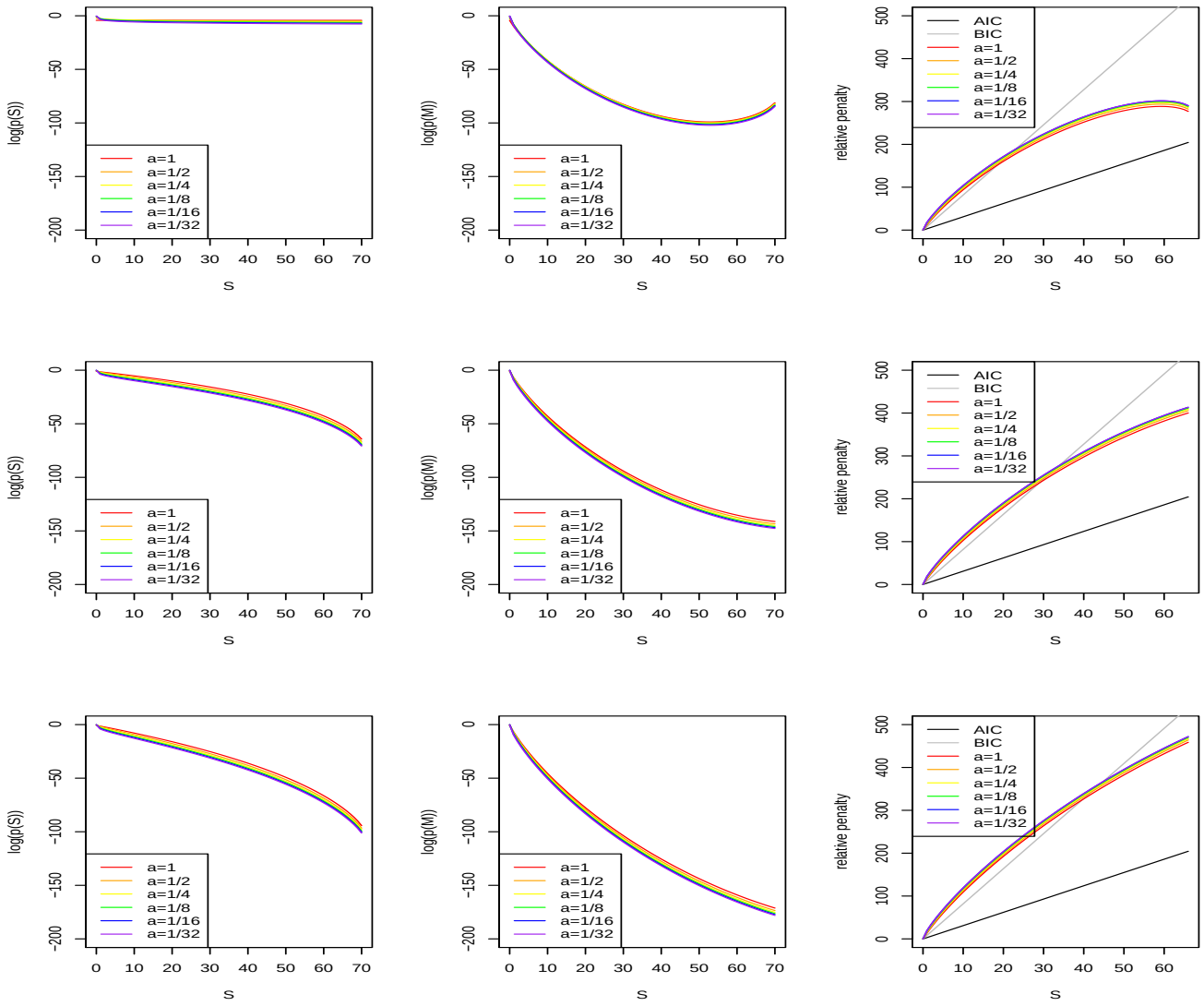

Figure 1: Plots of the prior on the model size $p(S)$ (column 1), the prior of the individual models, $p(\mathcal{M}_{\boldsymbol{\gamma}})$ (column 2), for each of the models of a given size $s_{\boldsymbol{\gamma}}$ and the relative penalties for each individual model for a given size $s_{\boldsymbol{\gamma}}$, $PEN(\mathcal{M}_{\boldsymbol{\gamma}})$ (column 3). In row 1 we assume $b = 1$, in row 2 $b = S/2$ and in row three $b = S$, where $S = 70$. For each plot, the color of the curve represents the value of a.

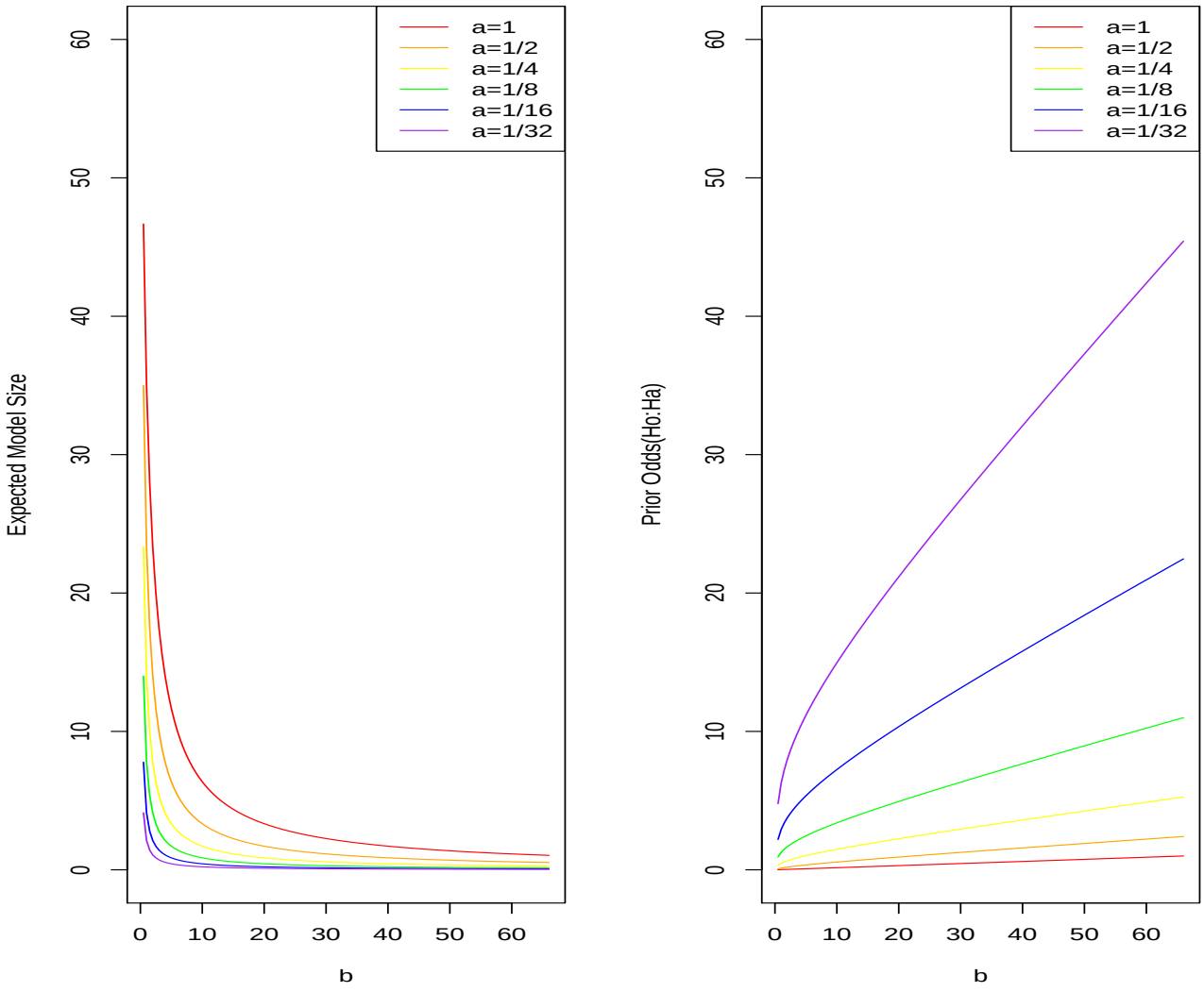

Figure 2: Plot of the expected model size and the prior odds of the null hypothesis to the alternative hypothesis, $\frac{p(H_0)}{p(H_A)}$, as functions of a and b and assuming $S = 70$.

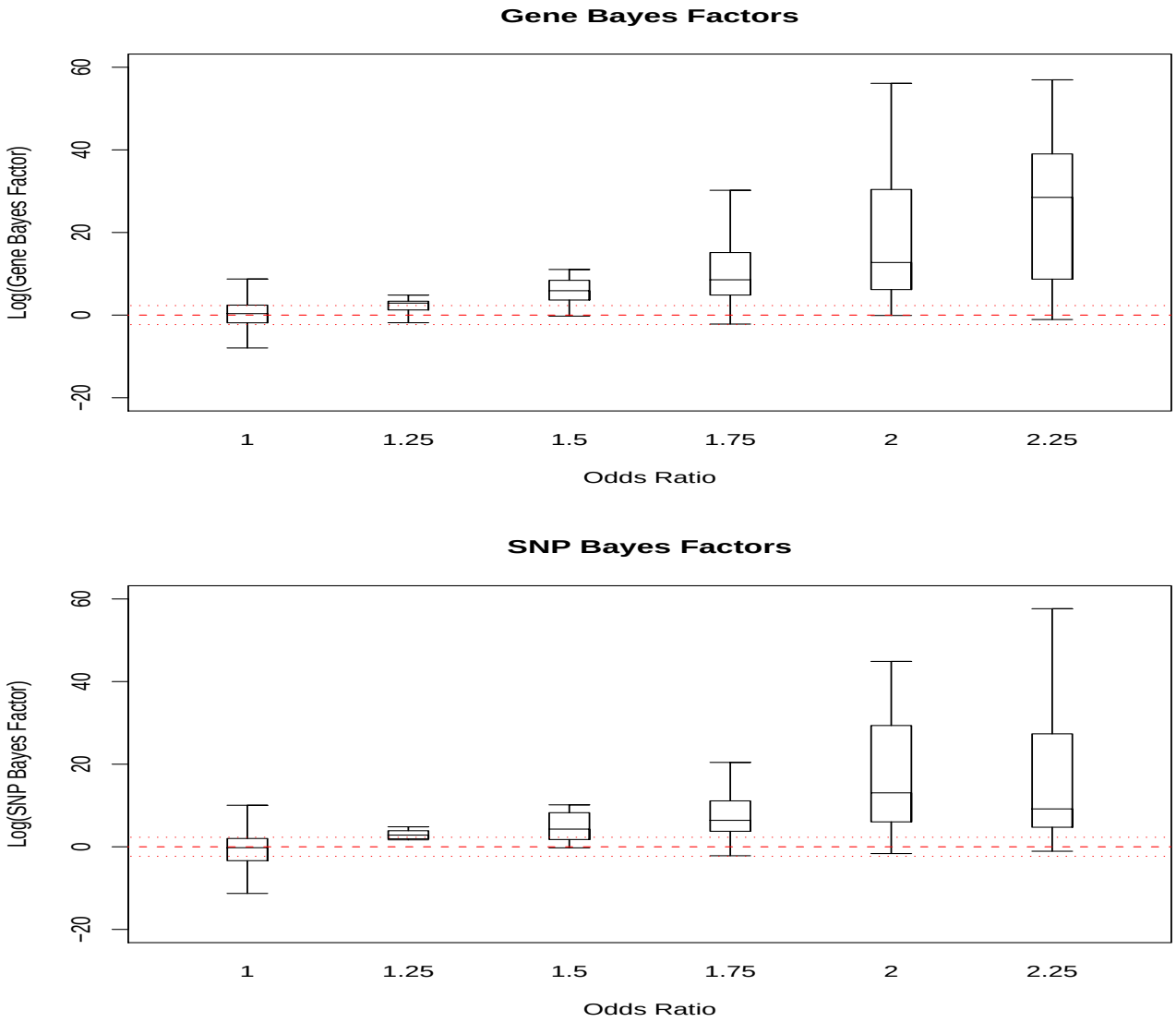

Figure 3: Boxplots of Log Gene and SNP Bayes factors respectively as a function of simulated ORs for Beta Binomial Prior with $b = S, a = \frac{1}{8}$. The dotted dash lines represent log(10), log(1) and log(−10) respectively. Gene and SNP false positive and true positive rates are given in the titles of the plots based on thresholding the Bayes factors at 10.

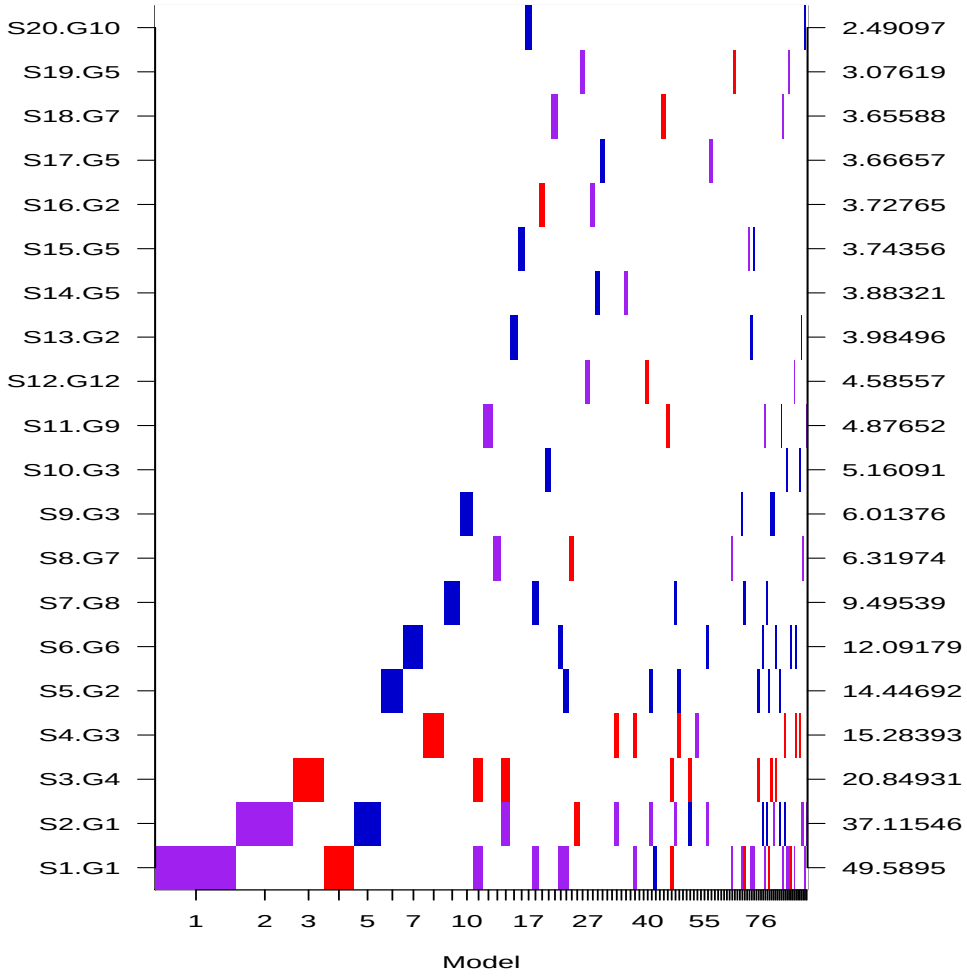

Figure 4: Image plot of the SNP inclusion indicators for the top 20 SNPs and the top 100 Models. The color of the inclusion block corresponds to the genetic parameterization of the SNP in that model. Purple corresponds to a log–additive parameterization, red to a dominant parameterization and blue to a recessive parameterization. SNPs are ordered on basis of their marginal SNP Bayes Factors which are plotted on the right axis across from the SNP of interest. Width of the column associated with a model is proportional to its estimated model probability.

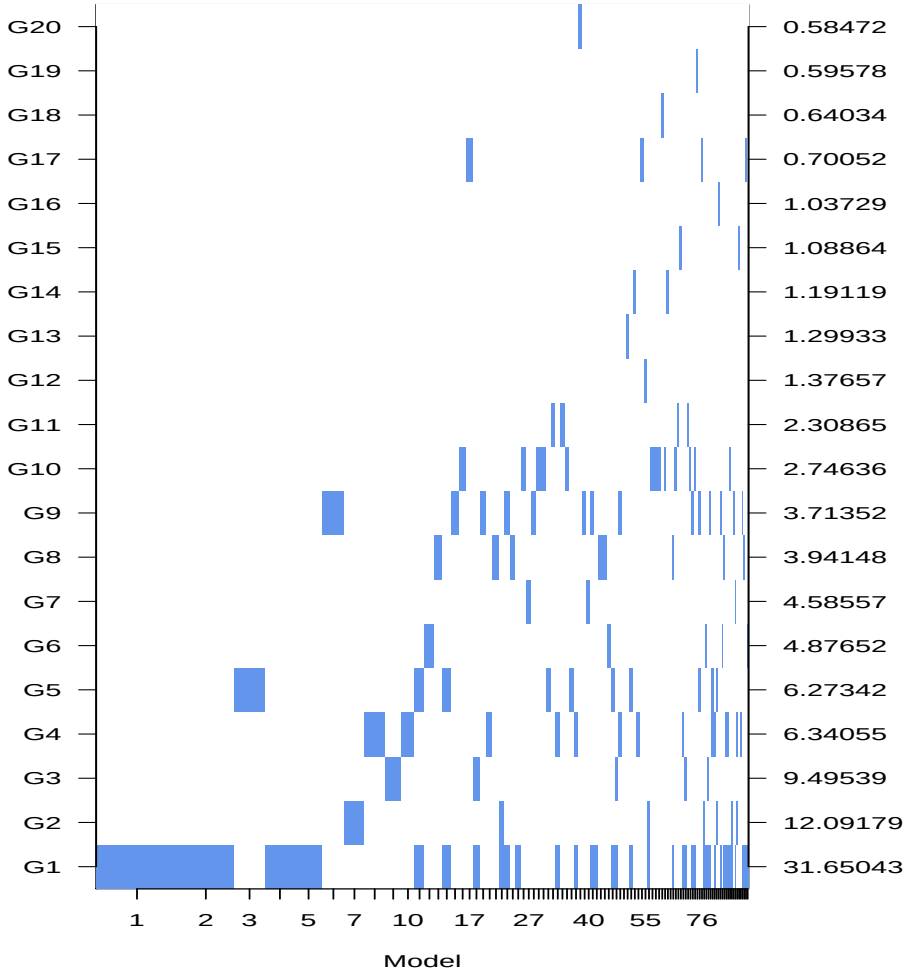

Figure 5: Image plot of the gene inclusion indicators for the top 100 Models. Genes are ordered based on their marginal gene Bayes Factors which are plotted on the right axis. Columns correspond to models and have width proportional to the estimated model probability, models are plotted in descending order of posterior support. The color is chosen to be neutral since the genetic parameterizations are not defined at the gene level.

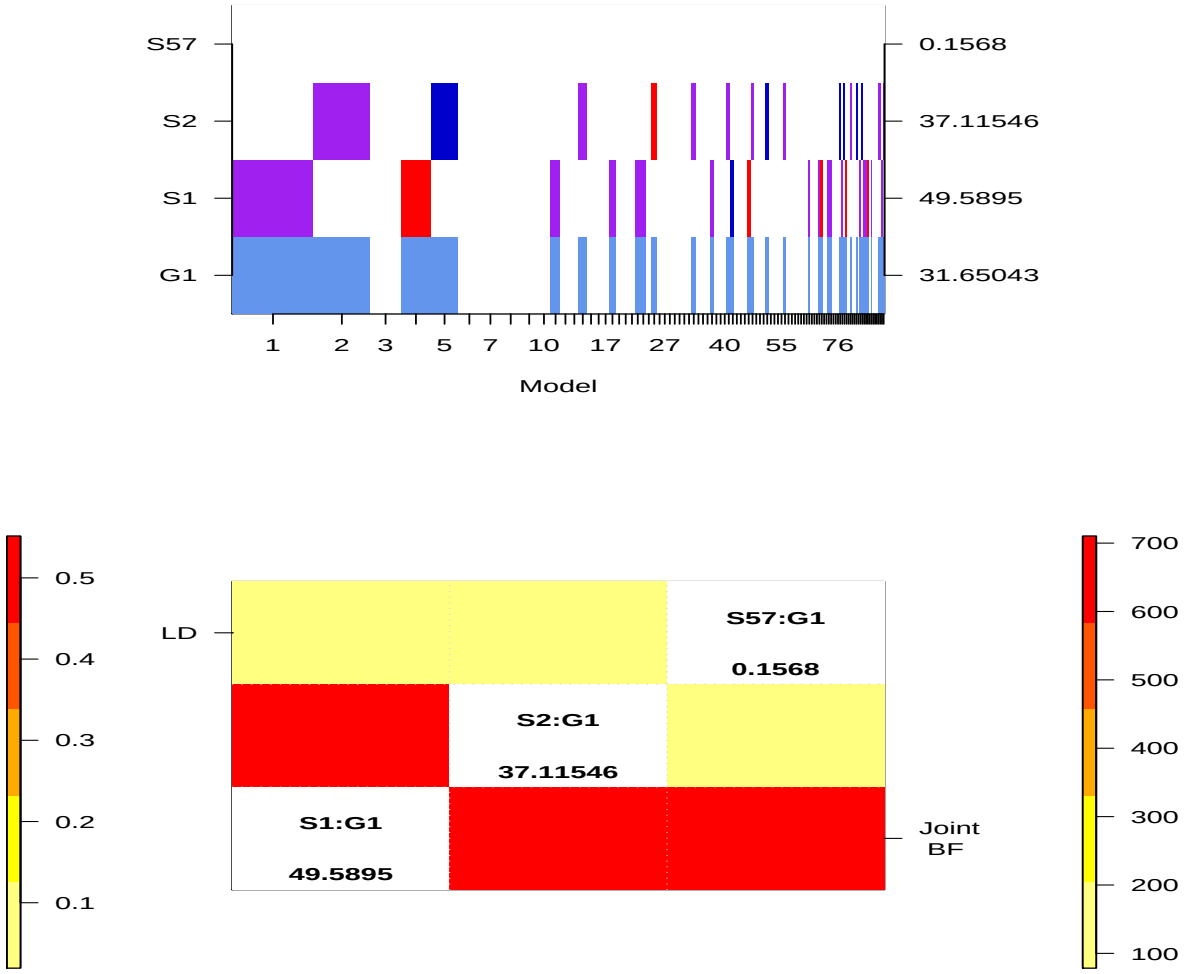

Figure 6: Upper panel: Image plot of the SNP and gene inclusions for G1 in the top 100 Models. Color scheme is as in the previous figures. Lower panel: Image plot of the pairwise LD (squared correlation) of the G1 SNPs (upper triangle of the plot) and the joint Bayes factors for each pair of SNPs (lower triangle of the plot). Marginal SNP Bayes factors are plotted on the right axis. Joint Bayes factors are defined as the ratio of the posterior odds that both of the SNPS appear in the model to the prior odds that both of the SNPs appear in the model.

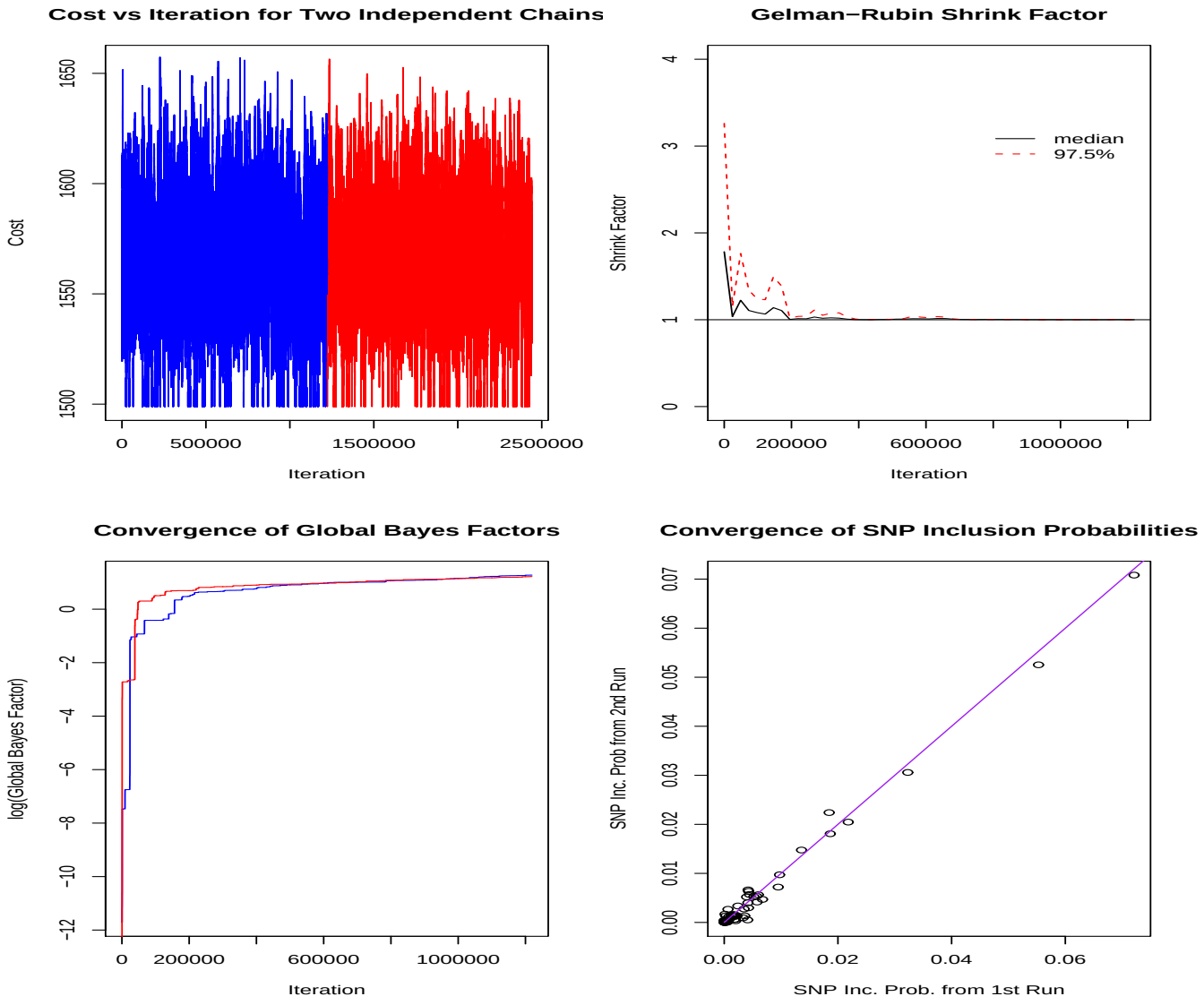

Figure 7: Upper left panel:Trace plot of the cost values of the models visited over each iteration for each of the two independent runs. Upper right panel: Plot of the Gelman-Rubin convergence diagnostic (see Gelman and Rubin (1992)) of the cost values of the models visited in the two independent chains. Lower left panel: Plot of the global Bayes factor computed across iterations for each independent chain. Lower right panel: Plot of the SNP inclusion probabilities for one of the independent runs vs. another independent run.